John von Neumann Institute for Computing

Sub-type Specific Sites for SMAD Receptor Binding Identified by Sequence Comparison Using "Sequence Harmony"

K. Anton Feenstra, Walter Pirovano, Jaap Heringa

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 73-78, 2006.

http://www.fz-juelich.de/nic-series/volume34

Sub-type Specific Sites for SMAD Receptor Binding Identified by Sequence Comparison Using "Sequence Harmony"

K. Anton Feenstra*, Walter Pirovano*, and Jaap Heringa

Centre for Integrative Bioinformatics VU (IBIVU)
Free University, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands
E-mail: heringa@few.vu.nl

We have examined the sub-type specific functions within the MH2 domain of the SMAD-family of transcription factors and found that our novel algorithm, called "Sequence Harmony", has a high specificity for identification of sites important for the functional differences. For the SMAD MH2 domain, 40 sub-type specific functional sites are predicted, which in the structure form clusters of similar function, *i.e.* for receptor binding, co-repressor binding and binding to transcription factors. From these clusters, putative functions were assigned to eleven out of fourteen predicted functional sites with unknown function. We propose these fourteen sites of unknown function as interesting candidates for further (experimental) investigation.

1 Introduction

Protein families and sub-families separated on the basis of functional properties.[1,2] It is therefore not surprising that a fair number of methods are in use for the comparison of amino-acid composition at different positions between groups of proteins from different families and/or sub-types.[3,4] It is surprising, however, that apparently in the vast majority of those studies, relatively little thought has been given to the underlying formalism of sequence comparison. Starting from a multiple sequence alignment (MSA) of the proteins of interest, the aim is to identify sites that are possibly conserved within a group, but certainly different between the groups.

Current practice seems to focus on sites that are conserved in both groups, but still different between them,[3] thereby excluding sites that are not totally conserved but nevertheless different between groups. This may not seem a serious problem at first hand, but let us consider an example of proteins that bind a certain receptor (the 'binders') and those that don't (the 'non-binders'). Certainly, one can expect sites that are crucial for binding to be conserved in the group of 'binders'. To exclude binding, on the other hand, several of many reasons may suffice and it seems imprudent to expect those sites to be conserved as well in the group of 'non-binders'. If the binders also interact with different, related, receptors, also the restriction to sites conserved in the group of 'binders' may not be a sensible one. Relative entropy is a measure for the difference in information content between both distributions of amino acid types.[5] Unfortunately, for those sites that interest us the relative entropy is degenerate. We will introduce an alternative similarity measure named *Sequence Harmony* for comparison of groups of sequences.

We have applied our method to the interactions of the SMAD proteins with the cell-membrane associated receptors TBβRI and BMPRI. The SMAD-TBβRI and SMAD-BMPRI interactions are relatively well-studied[6] and provide a good background of experimental data on which to validate our method.

** KAF and WP acknowledge financial support from Biorange/Netherlands Bioinformatics Centre.*

2 Theory

Commonly used methods for sequence comparison from multiple sequence alignments share several drawbacks. As an alternative solution, we propose:

$$SH_i^{AB} = \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B} \quad (1)$$

where the 'relative entropy' of group A is calculated relative to the sum of the probabilities of both groups ($p^A + p^B$). This function operates opposite to the relative entropy; zero for maximally different sites and one for sites with identical distributions. It is therefore that we coin this measure *Sequence Harmony* (*SH*) as it indicates the amount of correspondence of amino acid composition between two groups of sequences.

3 Results

Protein sequences for the R-SMADs were collected using the NCBI query for sequence retrieval (www.ncbi.nih.gov), yielding 32 sequences: 17 BR-SMADs and 15 AR-SMADs, including that of the SMAD2-MH2 structure (PDB-ID: 1KHX[7]), and aligned using PSI-Praline (www.ibivu.cs.vu.nl/programs/pralinewww).[8] From the alignment obtained, the MH2 domain was selected for further analysis, and divided according into two subgroups, AR-SMADs (binding TBβRI) and BR-SMADs (binding BMPRI).

Sequence Harmony was calculated using Eq. 1 for AR-SMADs and BR-SMADs at all positions in the alignment, see Figure 1. Relatively few sites are completely non-harmonious (32 *SH*=0) whereas the vast majority are overall conserved (137 *SH*=1), and the remaining are intermediate (44 0<*SH*<1). Furthermore, the low harmony sites are not spread completely randomly along the sequence, but clusters can be seen of up to about five low harmony sites spread over a sequence stretch of up to about ten residues.

Out of the 40 low harmony sites, 26 have a known function (65%), and for the 32 non-harmonious sites, 22 have a known function (69%). Of the 171 remaining high-harmony

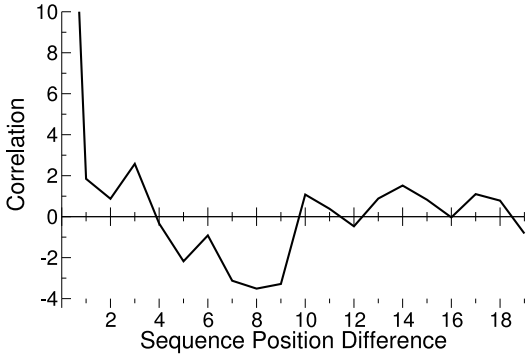

Figure 1. Sequence harmony for AR *vs.* BR-SMADs plotted along the sequence of the MH2 domain of SMADs.

sites, on the other hand, to the best of our knowledge no other specific sites have been identified as important for receptor binding specificity.

Figure 2 shows the SMAD2-MH2 domain colored by Sequence Harmony. The distribution of non-harmonious sites (red) is far from random, and in addition, the low harmony sites (orange) cluster along with them. In fact, we can identify a limited number of clustered regions high in low harmony sites, which are summarized in Table 1. The sites of known function in most clusters, allow us to assign putative functions to the remaining sites of unknown function.

Three of these clusters (#4, 5 and 7) are associated with receptor binding. The second-largest cluster (#2) is associated with c-Ski/SnoN interactions, and consists of a group of six largely non-sequentially low harmony sites. In addition, all sidechains point in the same general direction, forming a putative interaction surface. Three clusters (#1, 3 and 6) are associated with FAST1, Mixer and/or SARA binding. FAST1, Mixer share a SMAD interaction motif which is similar to that of SARA and have been shown to compete for binding to AR-SMADs.[10] Here as well, the sidechains form a putative interaction surface. The surface-patch formed by the sidechains of the largest cluster, #1, is less regular. The two smallest clusters (#8 and 9) are formed of sites of unknown function. In the structure of the functional trimeric form of the SMAD2-MH2 domain residues of cluster #8 are close to several sites of known function *across the protein interface*.

4 Discussion & Conclusion

Sites of low Sequence Harmony correspond very specifically to functionally relevant sites in the SMAD-MH2 domain, with a very sharp separation between conserved positions

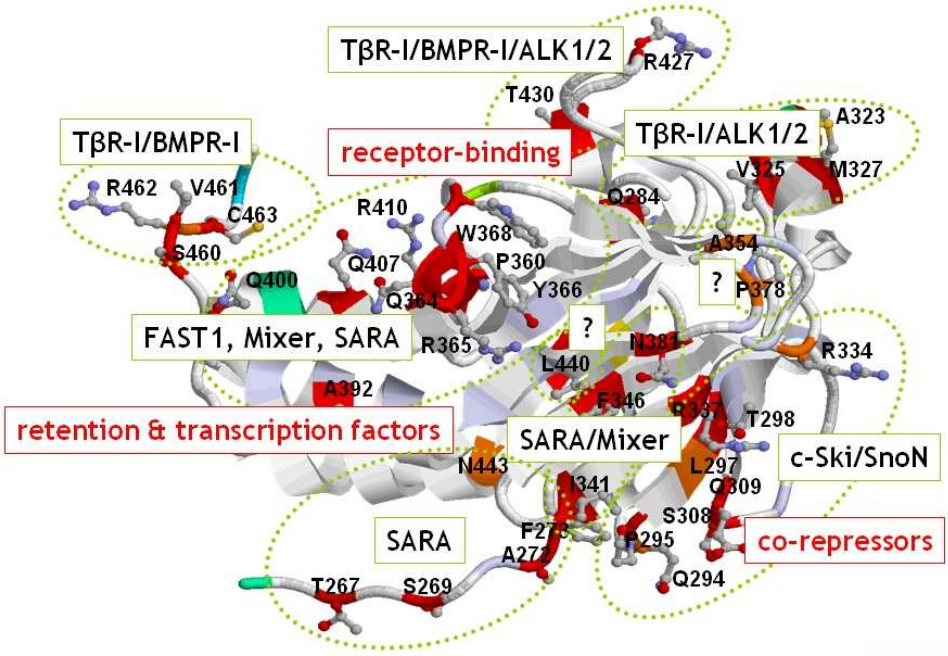

Figure 2. Sequence Harmony for AR-SMADs versus BR-SMADs colour-coded onto a SMAD2-MH2 structure (1KHX), *cf.* the values in Figure 1. Non-harmonious (*SH* zero) is red and labeled with residue names and numbers, maximal harmony (*SH* one) is white, intermediate values are rainbow colours from red to light blue.

Cluster	Sites	Ref.	Function
1	360, 364-366, 368, 392, 400, 407, 410	[7,11,10,12]	FAST1, Mixer, SARA, ?
2	294, 295, 297-298, 308, 309, 334, 337	[7,13]	c-Ski/SnoN, ?
3	267, 269, 272, 273, 443	[7]	SARA, ?
4	284, 323-325, 327	[6,14]	TβRI/BMPRI/ALK1/2
5	460-463	[12]	TβRI/BMPRI/ALK1/2
6	341, 346, 381	[7,10]	SARA/Mixer
7	427,430	[12]	TβRI/BMPRI/ALK1/2
8	354,378	-	?
9	440	-	?

Table 1. Structural clusters of low harmony sites (*cf.* Figure 2) and their respective known functions.

(which are the majority) and those that show a clear difference between the TGF-β and the BMP-binding sub-types. The Sequence Harmony scale corresponds to an intuitive interpretation of the differences in sequence composition and can be interpreted as sites that are more or less likely to be of functional importance. From the available experimental evidence in literature, it is difficult to identify false positive predictions, and almost no direct evidence is present to discriminate true from false negatives.

We have identified 13 sites of low Sequence Harmony in the SMAD-MH2 domain of unverified function and hereby suggest these as promising candidates for further elucidation of their function in determining the specificity of the TGF-β and BMP signalling pathways. Specifically, it would be very interesting to confirm (or rebuke) the putative functional roles we assigned to these sites of unknown function based on their proximity to low harmony sites of known function.

References

1. K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington. Homstrad: a database of protein structure alignments for homologous families. *Protein Sci.*, 7(11):2469–71, 1998.
2. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Res.*, 32(Database issue):D138–41, 2004.
3. S. S. Hannenhalli and R. B. Russell. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Molec. Biol.*, 303(1):61–76, 2000.
4. F. Pazos and M. J. E. Sternberg. Automated prediction of protein function and detection of functional sites from structure. *Proc. Nat. Acad. Sci. USA*, 101(41):14754–14759, 2004.
5. I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Molec. Biol.*, 336(5):1265–1282, 2004.
6. M. Huse, T. W. Muir, L. Xu, Y. G. Chen, J. Kuriyan, and J. Massague. The tgf beta receptor activation process: an inhibitor- to substrate-binding switch. *Mol. Cell*, 8(3):671–82, 2001.
7. G. Wu, Y. G. Chen, B. Ozdamar, C. A. Gyuricza, P. A. Chong, J. L. Wrana, J. Massague, and Y. Shi. Structural basis of smad2 recognition by the smad anchor for receptor activation. *Science*, 287(5450):92–7, 2000.

8. V. A. Simossis and J. Heringa. Praline: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic. Acids. Res.*, 33(Web Server issue):W289–94, 2005.
9. J. W. Wu, A. R. Krawitz, J. Chai, W. Li, F. Zhang, K. Luo, and Y. Shi. Structural mechanism of smad4 recognition by the nuclear oncoprotein ski: insights on ski-mediated repression of tgf-beta signaling. *Cell*, 111(3):357–67, 2002.
10. R. A. Randall, S. Germain, G. J. Inman, P. A. Bates, and C. S. Hill. Different smad2 partners bind a common hydrophobic pocket in smad2 via a defined proline-rich motif. *Embo J.*, 21(1-2):145–56, 2002.
11. Y. G. Chen, A. Hata, R. S. Lo, D. Wotton, Y. Shi, N. Pavletich, and J. Massague. Determinants of specificity in tgf-beta signal transduction. *Genes Dev.*, 12(14):2144–52, 1998.
12. R. S. Lo, Y. G. Chen, Y. Shi, N. P. Pavletich, and J. Massague. The l3 loop: a structural motif determining specific interactions between smad proteins and tgf-beta receptors. *Embo J.*, 17(4):996–1005, 1998.
13. M. Mizuide, T. Hara, T. Furuya, M. Takeda, K. Kusanagi, Y. Inada, M. Mori, T. Imamura, K. Miyazawa, and K. Miyazono. Two short segments of smad3 are important for specific interaction of smad3 with c-ski and snon. *J. Biol. Chem.*, 278(1):531–6, 2003.
14. Y. G. Chen and J. Massague. Smad1 recognition and activation by the alk1 group of transforming growth factor-beta family receptors. *J. Biol. Chem.*, 274(6):3672–7, 1999.